UNIVERSITÄT BERN

Hacking, Picking, Harking and Co.

Lessons from the Replication Crisis in Psychology

Boris Mayer, Cognitive Psychology, Perception & Research Methods Group

February 4, 2019, Interfaculty Research Cooperation – Decoding Sleep, Retreat

Unfolding of Psychology‘s Replication Crisis

2011: Bem‘s *Feeling the Future* in *JPSP*

Exemplary Result:

Participants better able to recall words that they were ***later*** randomly assigned to rehearse

How could such a well-respected scientist have amassed a large body of evidence for an obviously false hypothesis?

Journal of Personality and Social Psychology
2011, Vol. 100, No. 3, 407–425

Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem
Cornell University

Unfolding of Psychology‘s Replication Crisis

2011: Stapel‘s Fraud

News

Report finds massive fraud at Dutch universities

Investigation claims dozens of social-psychology papers contain faked data.

Published online 1 November 2011 | *Nature* **479**, 15 (2011) | doi:10.1038/479015a

Extreme result of the „strategic game hypothesis“? (Gigerenzer, 2018)

Gold Standard | Questionable Research Practices | Fraud

Unfolding of Psychology‘s Replication Crisis

2011: Stapel‘s Fraud

The Retraction Watch Leaderboard

Who has the most retractions? Here’s our unofficial list (see notes on methodology), which we’ll update as more information comes to light:

1. Yoshitaka Fujii (total retractions: 183) See also: Final report of investigating committee, our reporting, additional coverage
2. Joachim Boldt (96) See also: Editors-in-chief statement, our coverage
3. Diederik Stapel (58) See also: our coverage
4. Yoshihiro Sato (51) See also: our coverage
5. Yuhji Saitoh (49) See also: our coverage
6. Adrian Maxim (48) See also: our coverage
7. Jun Iwamoto (45) See also: our coverage
8. Chen-Yuan (Peter) Chen (43) See also: SAGE, our coverage
9. Fazlul Sarkar (41) See also: our coverage
10. Hua Zhong (41) See also: journal notice
11. Shigeaki Kato (40) See also: our coverage
12. James Hunton (37) See also: our coverage
13. Hyung-In Moon (35) See also: our coverage
14. Naoki Mori (32) See also: our coverage
15. Jan Hendrik Schön (32) See also: our coverage

retractionwatch.com/the-retraction-watch-leaderboard/

Unfolding of Psychology‘s Replication Crisis

2011: Simmons, Nelson & Simonsohn

Old paradox

- Overwhelming majority of published findings significant
- Overwhelming majority of studies underpowered

Old explanation: File drawer

- Claim: most failed studies not missing, but published, masquerading as successes

  ⛔ p-hacking

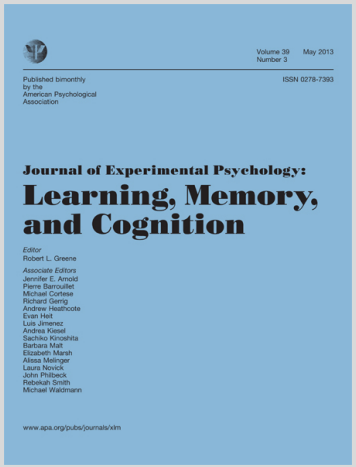
False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

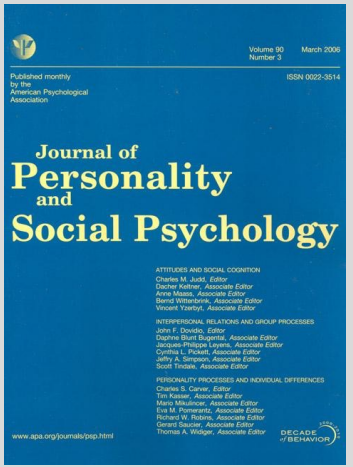
Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

Unfolding of Psychology‘s Replication Crisis

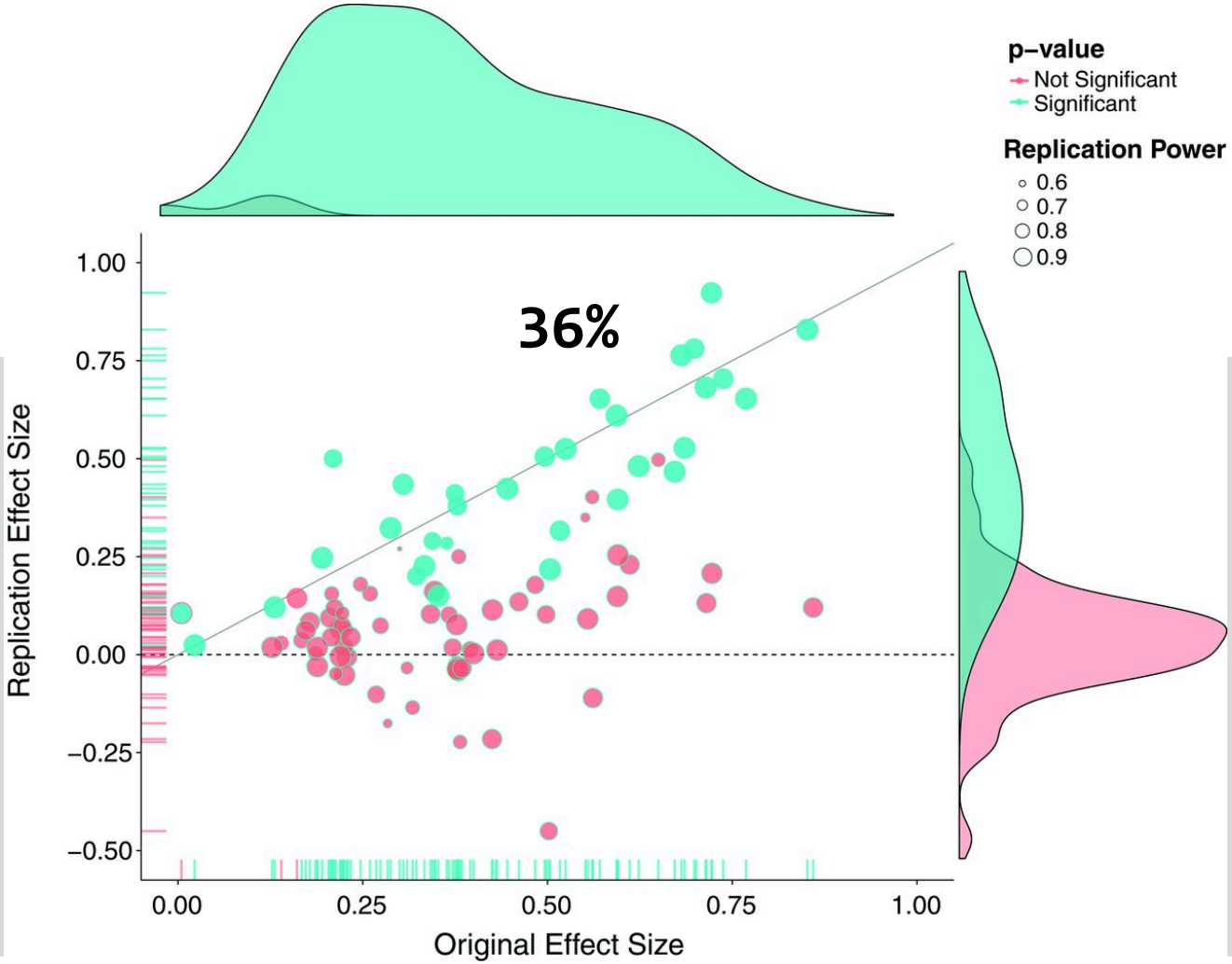
2012: Reproducibility Project: Psychology (RP:P)

- **Open Science Collaboration**
- **100 articles**
- **Median power = 95%**
- **Standardized protocol**
- **Feedback of original authors**





Open Sci. Collab. (2012, 2015)



Open Sci. Collab. (2015)

Unfolding of Psychology‘s Replication Crisis

2018: O‘Donell, Nelson et al.

The Relation Between Perception and Behavior, or How to Win a Game of Trivial Pursuit

Ap Dijksterhuis and Ad van Knippenberg
University of Nijmegen

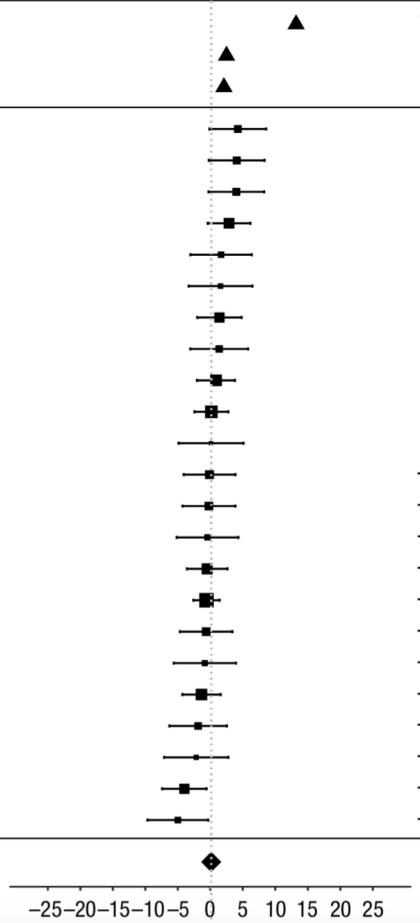
Registered Replication Report: Dijksterhuis and van Knippenberg (1998)

Perspectives on Psychological Science
2018, Vol. 13(2) 268–294

DOI: 10.1177/1745691618755704
www.psychologicalscience.org/PPS
SAGE

O‘Donell, Nelson et al. (2018)

Lab	Professor Condition Accuracy (%)	*n*	Hooligan Condition Accuracy (%)	*n*	Professor Condition – Hooligan Condition (% Difference)	Effect Size (%)	95% CI
Original Result	55.60		42.50			13.20	
Original Result (Replication 1)	51.40	77	48.90	101		2.50	
Original Result (Replication 2)	45.70	81	43.60	73		2.10	
Schulte-Mecklenbeck	58.47	50	54.22	60		4.24	[–0.12, 8.61]
Baskin	49.60	66	45.52	64		4.08	[–0.21, 8.36]
Braithwaite	61.01	66	57.00	70		4.01	[–0.28, 8.30]
Finnigan	52.68	137	49.79	140		2.89	[–0.39, 6.17]
Newell	54.48	64	52.82	58		1.66	[–3.05, 6.38]
O'Donnell	49.81	106	48.23	94		1.58	[–3.32, 6.49]
Tamayo	54.87	106	53.46	102		1.41	[–2.01, 4.83]
Karpinski	52.38	77	51.00	80		1.38	[–3.07, 5.83]
Klein	52.49	126	51.61	116		0.88	[–2.05, 3.81]
Keller	51.20	230	51.03	201		0.17	[–2.44, 2.79]
Shanks	53.29	72	53.17	63		0.11	[–4.90, 5.12]
Białobrzeska	52.71	64	52.82	72		–0.12	[–4.10, 3.87]
Koppel	58.04	63	58.25	76		–0.20	[–4.27, 3.87]
Philipp	52.38	70	52.81	70		–0.43	[–5.18, 4.33]
Ropovik	52.47	104	52.95	95		–0.48	[–3.62, 2.66]
Steele	50.94	267	51.52	277		–0.58	[–2.59, 1.43]
Susa	42.04	93	42.68	76		–0.63	[–4.68, 3.41]
Steffens	47.84	68	48.68	76		–0.84	[–5.63, 3.95]
Aczel	55.59	127	56.94	133		–1.35	[–4.30, 1.60]
Saunders	43.93	67	45.80	85		–1.87	[–6.31, 2.56]
McLatchie	53.21	53	55.37	54		–2.16	[–7.11, 2.78]
Aveyard	41.82	112	45.81	93		–3.99	[–7.42, –0.56]
Boot	47.92	64	52.91	86		–4.99	[–9.67, –0.31]
Meta-Analytic Average for Main Effect	51.70	2,252	51.49	2,241		0.14	[–0.71, 1.00]

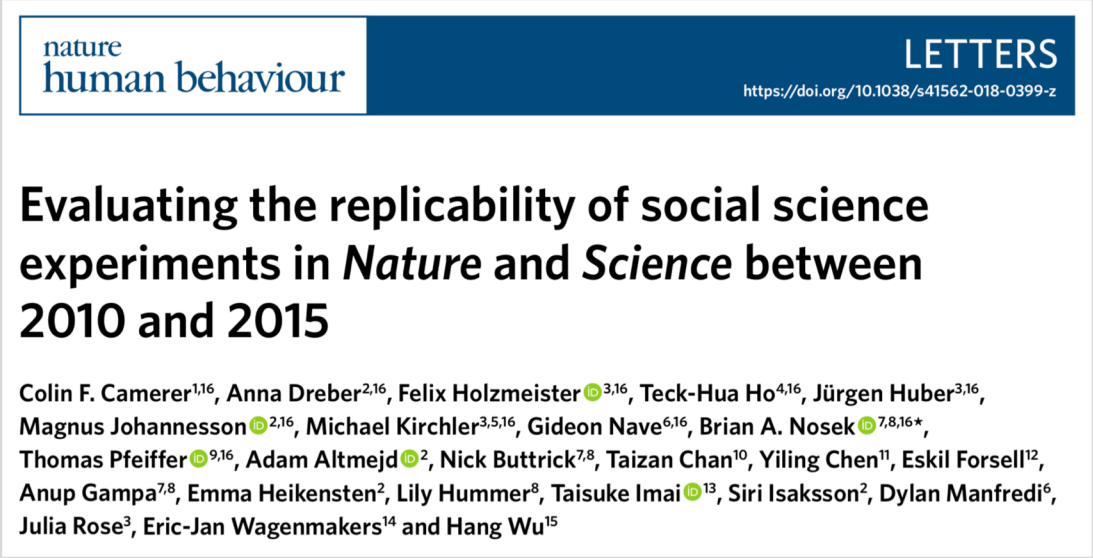


O'Donell, Nelson et al. (2018)

Unfolding of Psychology‘s Replication Crisis

2018: Camerer et al.

- Significant effect in same direction for 62% of studies
- Bayesian estimated true-positive rate 67%
- Relative effect size of true positives 71%, → both false positives and inflated effect sizes of true positives contribute to imperfect reproducibility



nature human behaviour

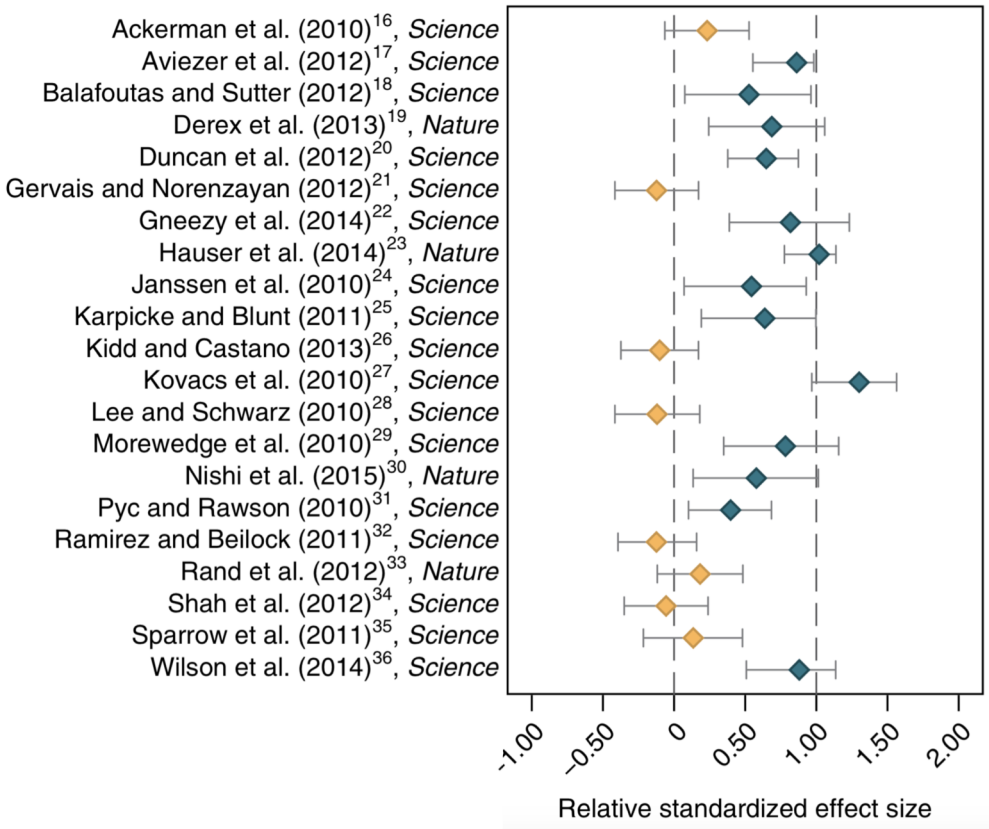
LETTERS

https://doi.org/10.1038/s41562-018-0399-z

Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

Colin F. Camerer[1,16], Anna Dreber[2,16], Felix Holzmeister[3,16], Teck-Hua Ho[4,16], Jürgen Huber[3,16], Magnus Johannesson[2,16], Michael Kirchler[3,5,16], Gideon Nave[6,16], Brian A. Nosek[7,8,16]*, Thomas Pfeiffer[9,16], Adam Altmejd[2], Nick Buttrick[7,8], Taizan Chan[10], Yiling Chen[11], Eskil Forsell[12], Anup Gampa[7,8], Emma Heikensten[2], Lily Hummer[8], Taisuke Imai[13], Siri Isaksson[2], Dylan Manfredi[6], Julia Rose[3], Eric-Jan Wagenmakers[14] and Hang Wu[15]

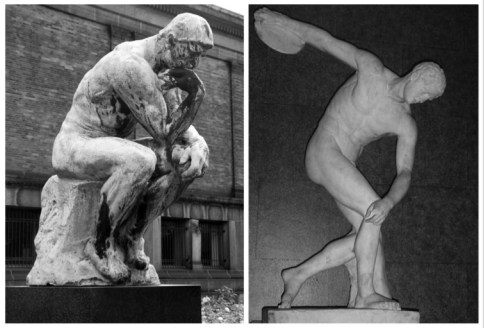
Unfolding of Psychology‘s Replication Crisis



Analytic Thinking Promotes Religious Disbelief

Will M. Gervais* and Ara Norenzayan*

Gervais & Norenzayan (2012)



«Given the present results […] we no longer have confidence in the finding that viewing pictures of *The Thinker* reduces self-reported religious belief.»

Gervais & Norenzayan (2018, *p.* 1)

Steile Thesen, nichts gewesen

Zweifelhafte Studien Kaum eine Disziplin produziert so viele Forschungsergebnisse, die einer Überprüfung nicht standhalten, wie die Psychologie. Steckt das Fach in der Krise – oder hat eine Selbstreinigung eingesetzt?

Sebastian Herrmann

Einer tröstenden Geschichte hört jeder gerne zu. Die Psychologin Amy Cuddy von der Harvard Business School veröffentlichte vor einigen Jahren eine Studie, die sich zu einem solchen Seelenbalsam kondensieren liess. Wer für eine Minute lang in einer Power-Pose verharre, fühle sich anschliessend tatsächlich mächtiger und verhalte sich risikofreudiger. Wer Dominanz ausstrahle, empfinde sich selbst als durchsetzungsfähig, so die frohe Botschaft.

Oh, wie schön, mag sich das Publikum gedacht haben, das klingt nach einem praktikablen Rezept: Ab sofort nehmen wir raumgreifende Körperhaltungen ein, und schon passt sich das Leben unseren Wünschen an. Eine Spitzenbotschaft, die Amy Cuddy auch in einem TED-Talk ausführen durfte, dessen Video etwa 50 Millionen Mal angesehen wurde.

Dummerweise liefert die Realität simple Botschaften stets mit einem Haken: Sie stimmen meist nicht. Auch im Fall der Power-Posen verwandelt die Körperhaltung alleine niemanden in ein Alphatier. Andere Forscher hatten in der Folge vielfach probiert, die Versuche zu wiederholen – es gelang ihnen nicht, die gleichen Ergebnisse zu erzielen.

Gut möglich, dass die Originalstudie aus dem Fachblatt «Psychological Science» ein falsch-positives Ergebnis erbracht hatte, also die Existenz eines Effekts nahegelegt hatte,

Mit Power-Pose zu mehr Durchsetzungsfähigkeit: Klingt gut, doch entsprechende Studienresultate konnten nie wiederholt werden. Foto: Chris Sorensen

Der Bund, Dec. 17, 2018

But earlier…

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key factors that influence this problem and some corollaries thereof.

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on p-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2×2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let R be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. R is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets $\text{PPV} = (1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

Abbreviation: PPV, positive predictive value

John P. A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

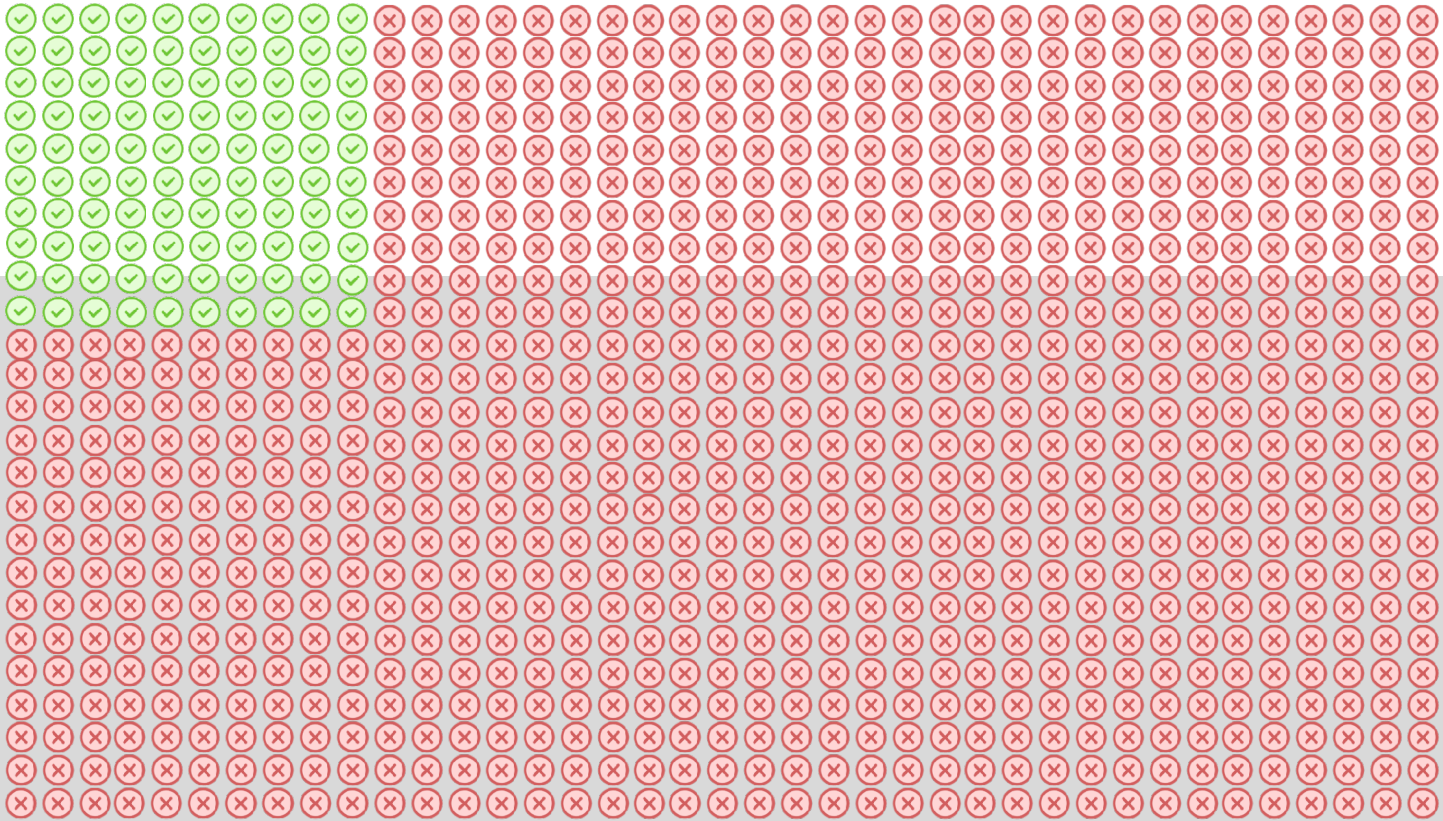
Competing interests: The author has declared that no competing interests exist.

DOI: 10.1371/journal.pmed.0020124

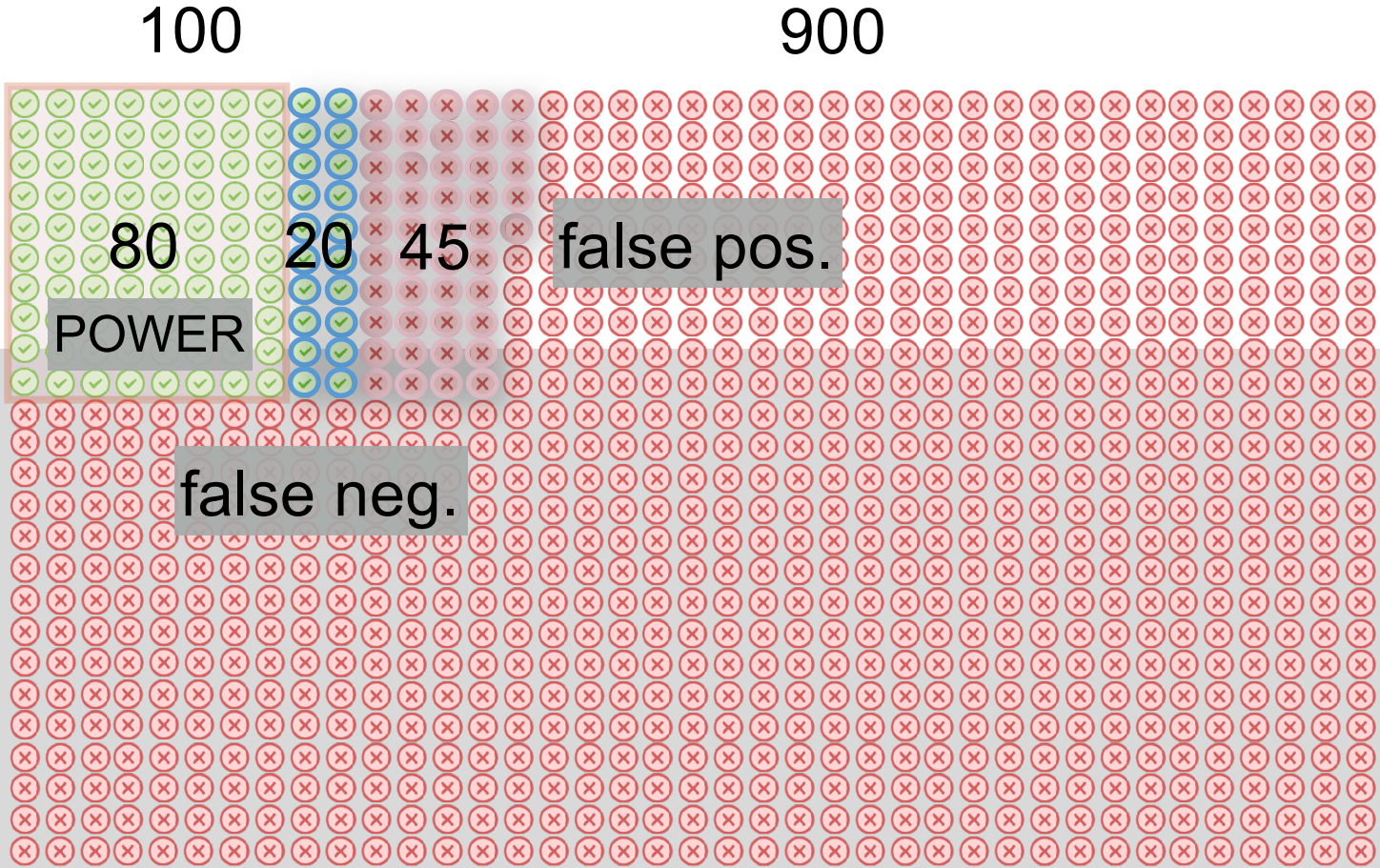
Ioannidis (2005)

100

900



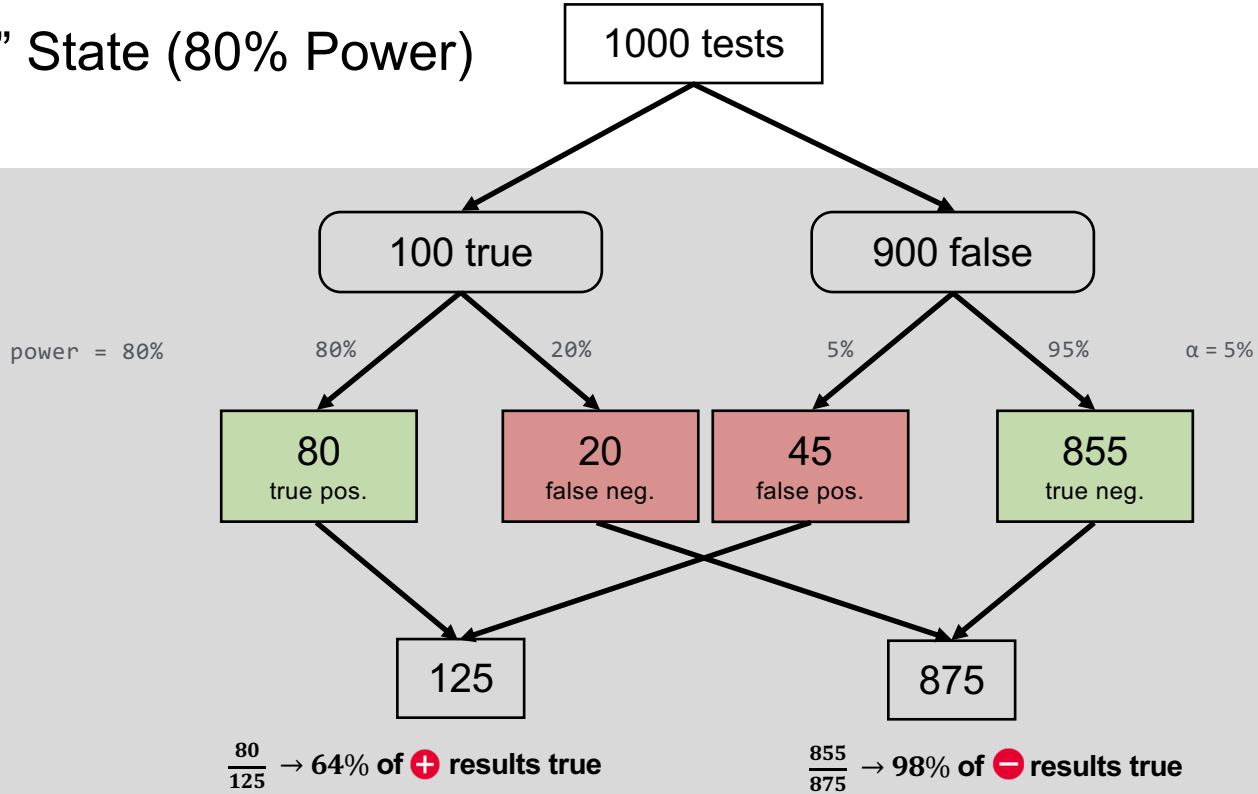
State of the world



UNIVERSITÄT BERN

QRP 1: Low Power

“Ideal” State (80% Power)



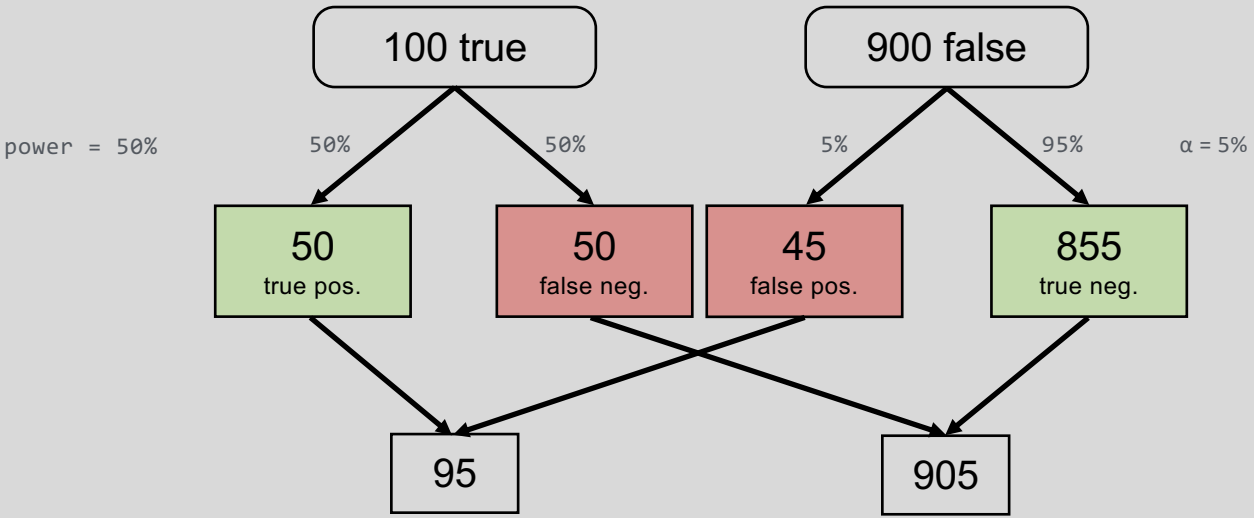
$\frac{80}{125}$ → 64% of ⊕ results true

$\frac{855}{875}$ → 98% of ⊖ results true

94% of published results true

QRP 1: Low Power

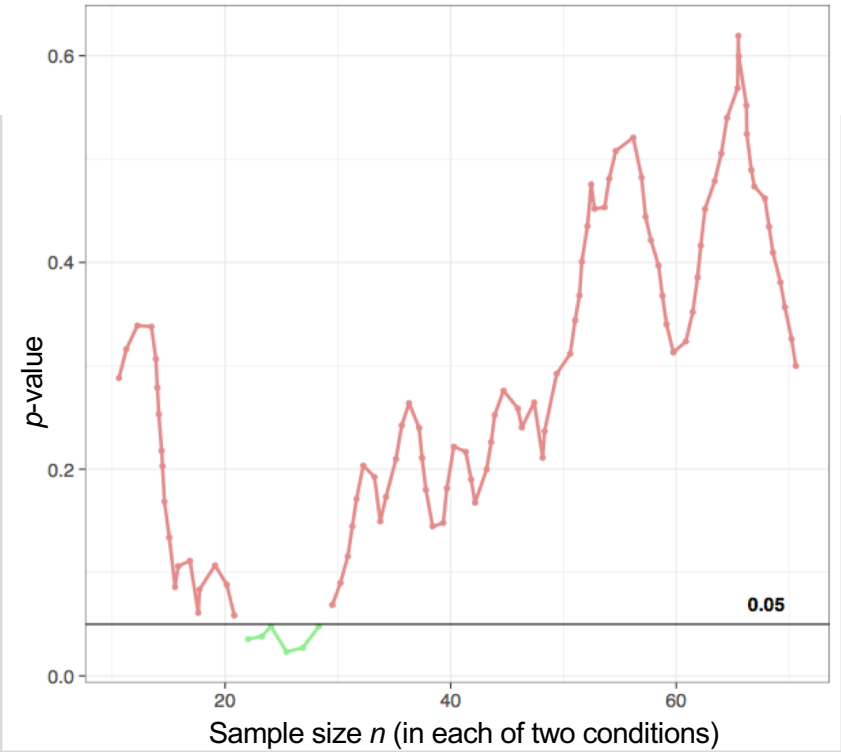
Actual State (~ 50% Power)



90% of published results true

QRP 2: “Mild” p-Hacking

Optional Stopping

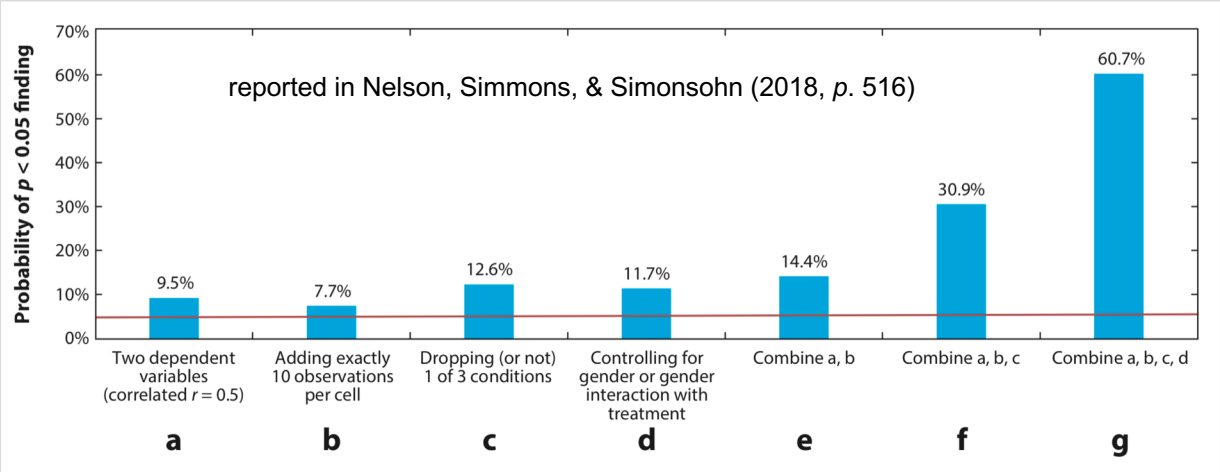


Simmons, Nelson, & Simonsohn (2011, *p*. 1362)

QRP 2: "Mild" p-Hacking

Simmons, Nelson, & Simonsohn (2011)

- Two-condition between-subjects design ($n = 20$ per cell) drawn from same normal distribution (no effect)
- 15,000 simulated studies
- *y* axis depicts share of studies for which **at least one** attempted analysis was significant.



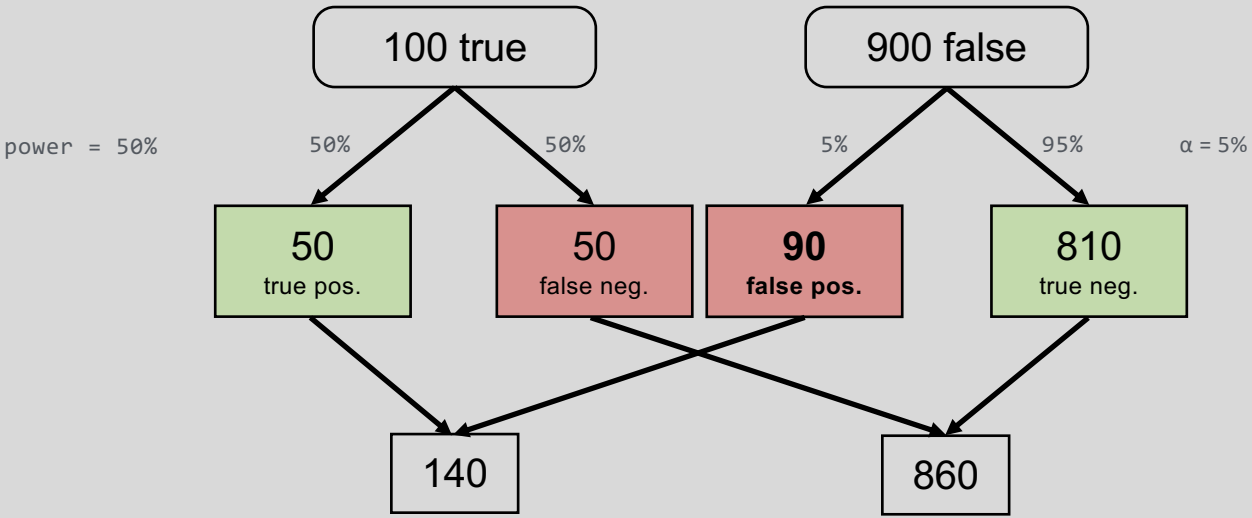
a: three t-tests, one on each of two dependent variables + third on the average of the variables

b: one t-test after collecting 20 observations per cell + another t-test after collecting an additional 10 observations per cell

c: t-tests for each of the three possible pairings of conditions + OLS regression for linear trend of all three conditions

d: t-test + 2-factorial ANOVA including gender main effect + one with gender interaction (sign. effects for condition or condition x gender interaction)

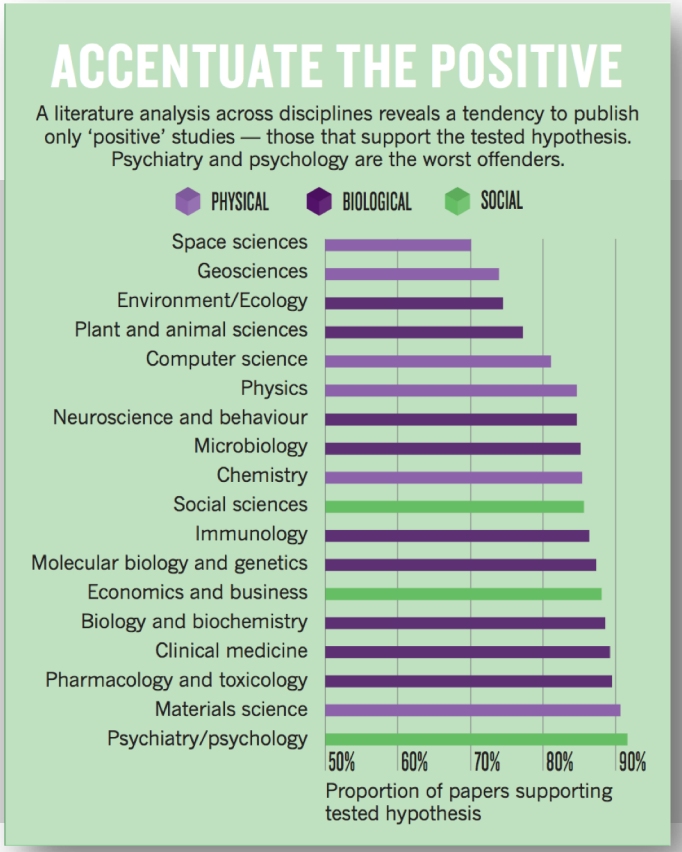
QRP 2: Optional Stopping & Mild p-Hacking (increasing type-I-error to $\alpha = 0{,}10$)



86% of published results true

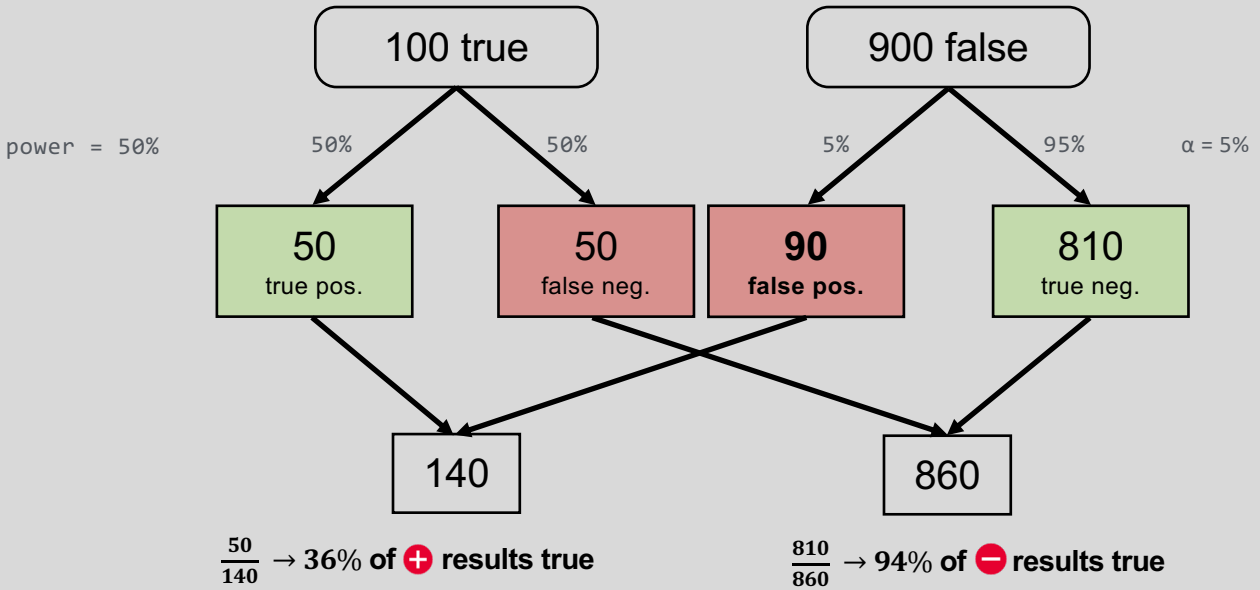
QRP 3: Cherry Picking and HARKing for Positive Results + File Drawer Problem

HARKing = Hypothesizing After Results Are Known



Yong (2012, *p*. 300)

QRP 3: Cherry Picking and HARKing for Positive Results + File Drawer Problem

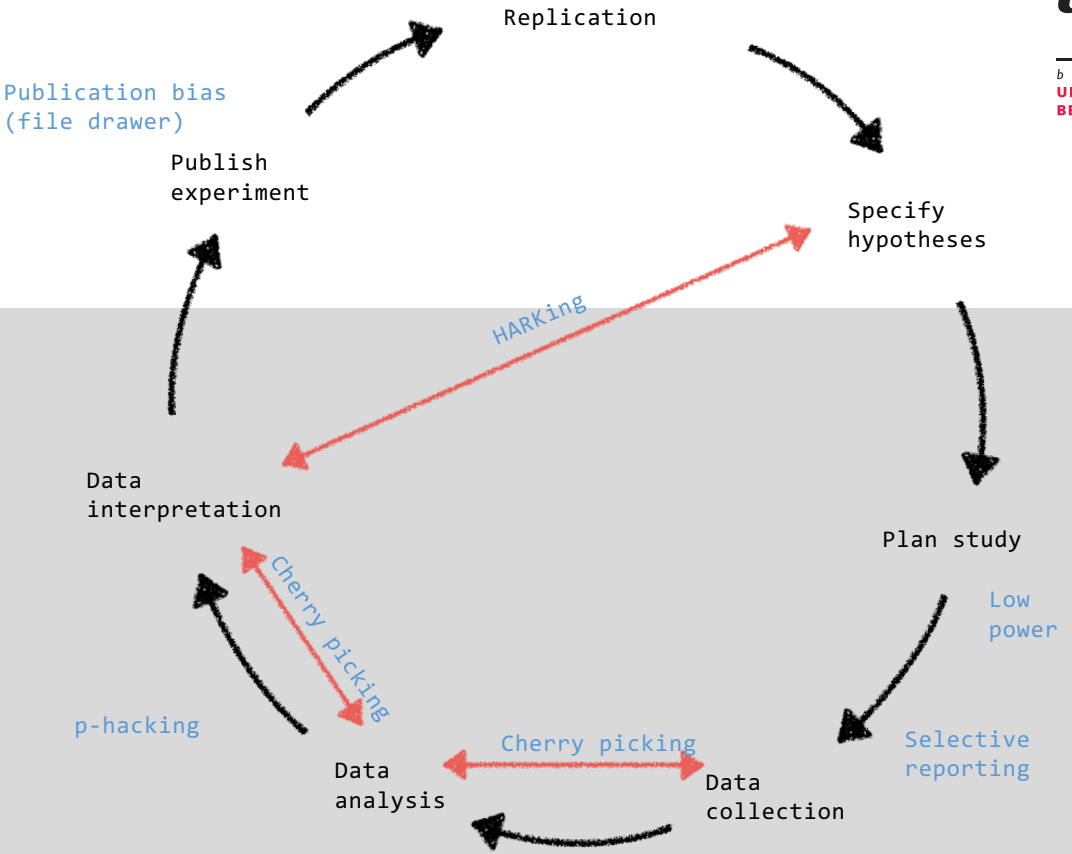


Assumption:
~85% of all published results positive and all positive results are being published, thus:

140 ⊕ (50 true)
25 ⊖ (23 true)

~44% of published results true (73 of 165)

QRP Summary


Replication
Publication bias
(file drawer)
Publish
experiment
Specify
hypotheses
HARKing
Data
interpretation
Plan study
Cherry picking
Low
power
p-hacking
Cherry picking
Selective
reporting
Data
analysis
Data
collection

What to do?

Research Integrity (Cumming, 2014)

Integrity of the research literature

- Meta-analytic (cumulative) approach
- Full reporting of all research that is «conducted to at least a reasonable standard»
- Close as well as conceptual replications combined with meta-analyses of existing research
- But: meta-analyses may exacerbate problem because of p-hacked studies (Nelson et al., 2018)

Integrity of researchers' values & behaviors

- Stop questionable research practices (incl. low power)
- Best way to avoid all biases is to «specify and commit to full details of a study in advance»
- Clear distinction between exploratory and confirmatory research
- Exploratory research can be pilot study like: «Discover how to prespecify a study that then must be reported»

What to do?

Open Science (Cumming, 2014; Nelson et al., 2018)

Preregistration and Documentation

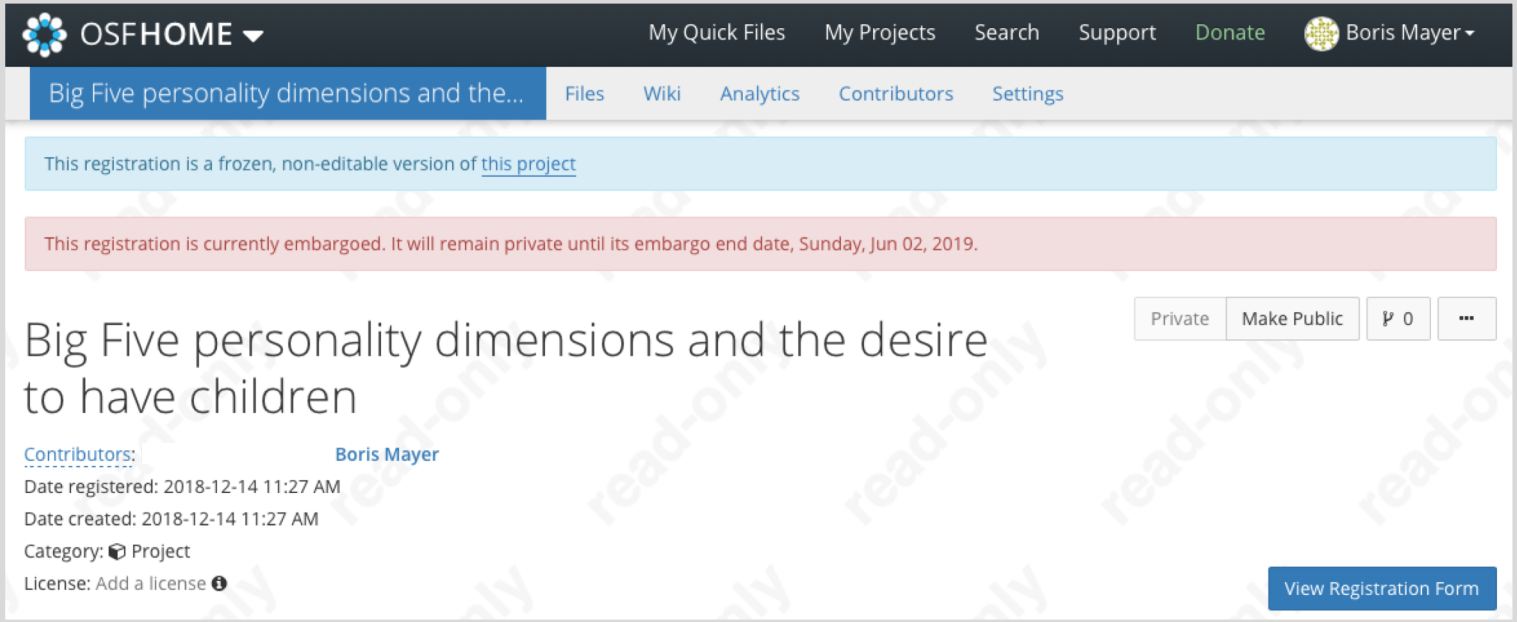
- Preregister all aspects of the planned study
- Open Data and Open Code
- Open Science Framework
  osf.io
- AsPredicted
  aspredicted.org
- AllTrials
  alltrials.net

What (exactly) is preregistration?

- Preregistration means specifying hypotheses, sample size, procedures and materials and data analytic methods
- Not every detail has to be specificed in advance
- Deviations possible but have to be declared
- Both exploratory and confirmatory research questions can be pre-registered
- Preregistration ≠ Registered Report

What to do?

Open Science

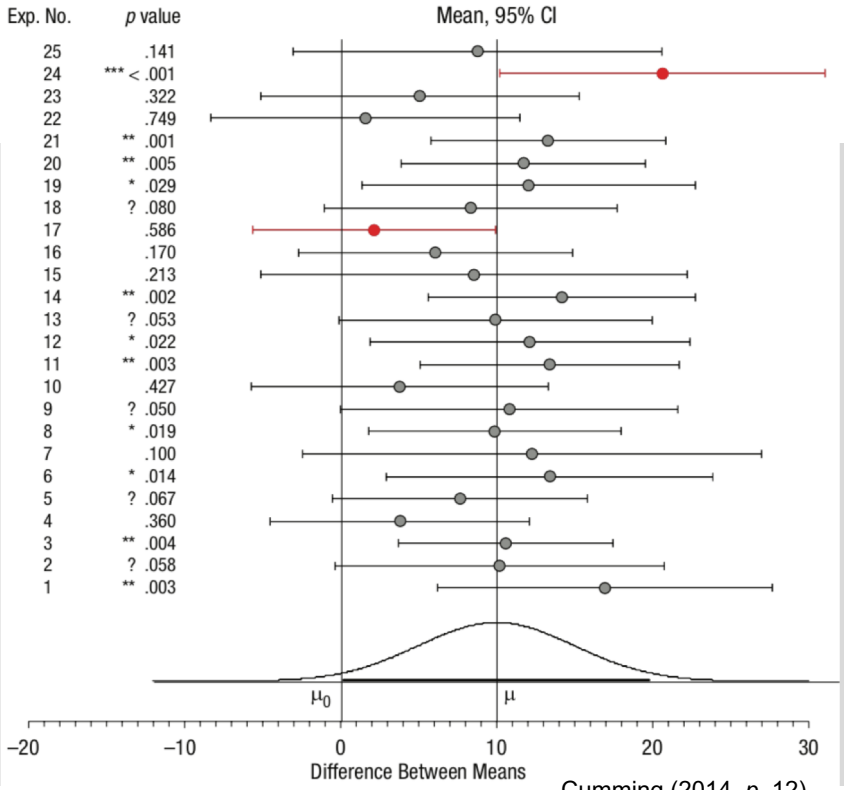


What to do?

Abandon p-values (1)

New Statistics (Cumming, 2014)

- Need to shift from reliance on NHST to estimation
- «NHST prompts us to see the world as black and white and to make our conclusions in dichotomous terms»
- Effect Sizes (ES) and Confidence Intervals (CI)
- **Example:** Two independent groups, each $n = 32$
- ES = 10 or Cohen's $\delta = 0.5$ (medium sized effect)
- p varies extremely (only 9 out of 24 «replicate»)
- 95%-CI = 83% prediction interval for replication (replication mean contained in 83% of CIs)
- CIs much more informative than p-values



Cumming (2014, *p.* 12)

What to do?

Abandon p-values (2)

Statistical Toolkit (Gigerenzer, 2018)

- *Strategic game hypothesis* vs. *Statistical ritual hypothesis*
- Researchers are stuck in the delusion that the «null ritual» answers research questions
- In fact, NHST as practiced today is not (even) in accordance with R. A. Fisher's null hypothesis test
- Psychology departments need to begin teaching statistical thinking and a more complete statistical toolkit, not rituals

ASA Statement on p-values (2016)

- p-values can indicate how incompatible the data are with a specified statistical model
- p-values do not measure the probability that the studied hypothesis is true
- Scientific conclusions should not be based only on whether a p-value passes a specific threshold.
- Statisticians can replace p-values with approaches "…that emphasize estimation over testing, such as confidence, credibility, or prediction intervals; …" (p. 132)

References

Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407-425.

Camerer, C. F. et al. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour,* 2, 637-644.

Cumming, G. (2014). The new statistics: why and how. *Psychological Science, 25*, 7-29.

Dijksterhuis, A., & Van Knippenberg, A. (1998). The relation between perception and behavior, or how to win a game of trivial pursuit. *Journal of Personality and Social Psychology, 74*, 865-877.

Gervais, W. M., & Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. Science, 336, 493-496.

Gervais, W. M., & Norenzayan, A. (2018). Analytic atheism revisited. *Nature Human Behaviour, online*. doi: 10.1038/s41562-018-0426-0

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science,1*, 198-218.

Greenland, S. et al. (2016). Statistical tests, *P* values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology, 31*, 337–350.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*, 696-701.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511-534.

O'Donnell, M., Nelson, L. D. et al. (2018). Registered Replication Report: Dijksterhuis and van Knippenberg (1998). *Perspectives on Psychological Science, 13*, 268-294.

Open Science Collaboration (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*, 657-660.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. Science, 349(6251):aac4716

Simmons, J., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359-1366.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician, 70*, 129-133.

Yong, E. (2012). Bad copy. *Nature, 485*, 298-300.

Thank you!

boris.mayer@psy.unibe.ch

Special thanks to:
Michael Schulte-Mecklenbeck & Andrew Ellis